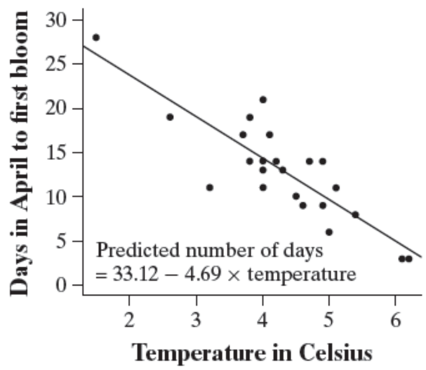
Chapter Review Exercises (page 202)

R3.1 (a) There is a moderate, positive linear association between gestation and life span. Without the outliers at the top and in the upper-right, the association appears moderately strong, positive and curved.
(b) The hippopotamus makes the correlation closer to 0 because it decreases the strength of what would otherwise be a moderately strong positive association. Because this point's x coordinate is very close to $\bar{x}$, it won't influence the slope very much. However, it makes the y intercept higher because its y coordinate is so large compared to the rest of the values. Because it has such a large residual, it increases the standard deviation of the residuals.
(c) Because the Asian elephant is in the positive, linear pattern formed by most of the data values, it will make the correlation closer to 1. Also, because the point is likely to be above the least-squares regression line, it will "pull up" the line on the right side, making the slope larger and the intercept smaller. Because this point is likely to have a small residual, it decreases the standard deviation of the residuals.

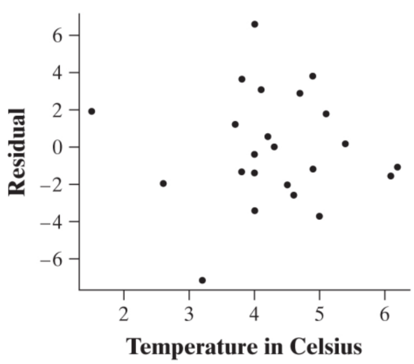
R3.2 (a) The slope is 0.0138 minutes per meter. For each increase of 1 meter in dive depth, the predicted duration increases by 0.0138 minutes.
(b) The y intercept suggests that a dive of no depth would last an average of 2.69 minutes; this obviously does not make any sense.
(c) When depth = 200, the predicted dive duration is $\hat{y} = 2.69 + 0.0138(200) = 5.45$ minutes.
(d) If the variables are reversed, the correlation will remain the same. However, the slope and y intercept will be different.

R3.3 (a) The least-squares regression line is $\hat{y}$ = 3704 + 12,188x where y represents the mileage of the cars and x represents the age.
(b) For a 6-year old car, the predicted mileage is $\hat{y}$ = 3704 + 12,188(6) = 76,832. The residual for this particular car is $y - \hat{y}$ = 65,000 – 76,832 = –11,832. In other words, this teacher has driven 11,832 fewer miles than predicted based on the age of the car.
(c) Because $r^2 = 0.837$ and the slope is positive, the correlation $r = +\sqrt{0.837} = 0.915$. This shows that there is a strong, positive linear association between the age of cars and their mileage.
(d) A linear model is appropriate for these data because there is no leftover pattern in the residual plot.
(e) s = 20,870.5: When using the least-squares regression line with x = car's age to predict y = number of miles it has been driven, we will typically be off by about 20,870.5 miles. r^2 = 83.7%: About 83.7% of the variability in mileage is accounted for by the linear model relating mileage to age.

R3.4 (a) The scatterplot is shown below. Average March temperature was chosen as the explanatory variable because changes in March temperature probably have an effect on the date of first bloom. Also, we are told that we want to predict the date of first bloom from the temperature.



(b) The correlation is $r = -0.85$. The least-squares regression equation is $\hat{y} = 33.12 - 4.69x$ where y represents the number of days and x represents the temperature. The correlation tells us that there is a strong, negative linear association between the average March temperature and the days in April until first bloom. The slope tells us that for every 1 degree increase in average March temperature, the predicted number of days in April until first bloom decreases by 4.69. The y intercept tells us that if the average March temperature was 0 degrees Celsius, the predicted number of days in April to first bloom is 33.12 (May 3). However, $x = 0$ is outside of the range of data, so this prediction is an extrapolation.
(c) No, $x = 8.2$ is well beyond the values of x we have in the data set. This prediction would be an extrapolation.
(d) The predicted number of days until 1st bloom is $\hat{y} = 33.12 - 4.69(4.5) = 12.015$. The residual is $y - \hat{y} = 10 - 12.015 = -2.015$. In this year, the actual date of first bloom occurred about 2 days earlier than predicted based on the average March temperature.
(e) The residual plot is given below. There is no leftover pattern in the residuals, indicating that a linear model is appropriate.



R3.5 (a) The slope of the regression line for predicting final-exam score from pre-exam totals is $b = 0.6\left(\frac{8}{30}\right) = 0.16$. The y intercept of the regression line is $a = 75 - 0.16(280) = 30.2$. Thus, the equation of the least-squares regression line is $\hat{y} = 30.2 + 0.16x$, where y = final exam score and x = total score before the final examination.
(b) Julie's predicted final exam score is $\hat{y} = 30.2 + 0.16(300) = 78.2$.
(c) Of all the lines that the professor could use to summarize the relationship between final exam score and total points before the final exam, the least-squares regression line is the one that has the smallest sum of squared residuals.
(d) Because $r^2 = 0.36$, only 36% of the variability in the final exam scores is accounted for by the linear model relating final exam scores to total score before the final exam. More than half (64%) of the variation in final exam scores is *not* accounted for by the least squares regression line, so Julie has a good reason to think this is not a good estimate.

R3.6 Even though there is a high correlation between number of calculators and math achievement, we shouldn't conclude that increasing the number of calculators will *cause* an increase in math achievement. It is possible that students who are more serious about school have better math achievement and also have more calculators.